The 2006 ISL Rich Transcription Speech-to-Text System

Christian Fügen, Matthias Wölfel, John McDonough, Shajith Ikbal, Florian Kraft, Kornel Laskowski, Mari Ostendorf, Tobias Gehrig, Sebastian Stüker, Ulrich Klee, Kenichi Kumatani

Institut für Theoretische Informatik
Universität Karlsruhe

May 3, 2006

IHM Condition: Acoustic Model Training Material

- As acoustic training material for the IHM condition we considered the corpora in Table :

ISL	ICSI	NIST	TED	CHIL	Hub4-BN
11h	72h	13h	13h	10h	180h

Table 1: Duration acoustic model training data.

- All corpora were sampled at 16 kHz with 16 bit sample resolution.
- All corpora were collected with high quality close-talking microphones except for the ISL corpus, which as collected with lapel microphones.
- Excluding the ISL and Hub4-BN data was found to yield a 1% reduction in WER.

Signal Processing

- All acoustic models trained for RT06s had a final feature length of 42 obtained by concatenating 15 consecutive frames, applying *linear discriminant analysis* (LDA), followed by a global *semi-tied covariance* (STC) transformation.
- The standard frontend was based on *Mel-frequency cepstral coefficents* (MFCCs) obtained with FFT analysis.
- A second frontend was based on a spectral envelope estimated with *minimum variance distortionless response* (MVDR) of order 30.
- No filter bank was used in the MVDR system, but the number of cepstral coefficients was increased from 13 to 20.
- Compared to the FFT, the MVDR frontend provides an increased resolution in low-frequency regions of the spectrum.
- The MVDR also provides detailed modeling of spectral peaks, but an approximation of spectral valleys; see [Wölfel2005].

Conventional Acoustic Model Training

- Acoustic model training was performed with fixed state alignments.
- Training was identical for MVDR and FFT systems, with 16,000 distributions over 4,000 codebook, with a maximum of 64 Gaussians per codebook.
- The complete training sequence was:
  1. linear discriminant analysis (LDA);
  2. first merge and split estimation of Gaussians;
  3. estimation of the global semi-tied covariance (STC) matrix;
  4. second merge and split estimation of Gaussians using LDA and STC matrices from prior steps.
- The second merge and split step provided an additional reduction in WER of 0.3% absolute.
- Since the 10hrs of CHIL training data were released by ELDA relatively late, MAP with a weight of 0.8 was used to obtain 0.6% reduction in WER.

Training Sequence Experiments

- WERs computed with first pass FFT systems with incremental VTLN and FSA estimation on the IHM condition of the NIST RT-06S development set.
- All decodings were done with a frame shift of 10msec.

Expt.	System	WER
A	ICSI+NIST+TED	34.8%
	+ CMU	35.1%
	+ BN97	36.0%
B	standard	32.3%
	second incr. growing	32.0%
C	w/o CHIL	32.1%
	with CHIL	31.5%

Table 2: Training setup experiments.

- All details can be found in [Fügen2006b].

Adapted Acoustic Model Training

- All adapted training iterations were based on Viterbi state alignments.
- Three additional iterations of maximum likelihood speaker-adapted training (ML-SAT) were applied to the FFT and MVDR models after MAP adaptation with the CHIL specific data.
- During ML-SAT, the CHIL data received a weight of 4.
- Both feature space adaptation (FSA) and maximum likelihood linear regression (MLLR) parameters were estimated during ML-SAT.
- The approximation described in [McDonough2002] was used to minimize disk space usage during ML-SATraining.

Alternate Phone Set Acoustic Model

- A third acoustic model was trained with the MVDR frontend and the PRONLEX phone set.
- Initial training and recognition lexicons were obtained by merging the Callhome English and LIMSI SI-284 dictionaries.
- Missing pronunciations were added with the Fisher grapheme-to-phoneme conversion tool.
- The PRONLEX model was initialized from a context indepedent system trained beginning from a global mean and covariance.
- The final context dependent system had 3,000 codebooks with a maximum of 64 Gaussians each, and 24,000 distributions.
- The final system was trained with only FSA during ML-SAT.

Language Model Training

- All language models were trained text and transcriptions from the following corpora:
  - subset of CHIL development data;
  - RT04s development and evaluation sets;
  - AMI, CMU, ICSI, and NIST meeting sets;
  - TED;
  - Hub4 broadcast news;
  - recent conference proceedings from 2002 to 2005;
  - web data collected at the University of Washingthon related to CMU, ICSI, and NIST meetings;
- All LMs were built using the SRILM toolkit.
- Chen and Goodman's modified Kneser-Ney technique was used for discounting.
- Pruning was performed after the interpolation of the LM components.

Harvesting Web Data for Language Model Training

- Query generation was based on:
  - topic phrases generated by computing bigram-based *term frequency inverse document frequencies* (tf-idfs) of the proceedings papers mentioned above;
  - all topic phrases with stop words were removed;
  - general tri- and 4-gram phrases extracted from the CHIL dev data—yielding the CHILweb set—and the CMU, ICSI, and NIST meetings—yielding the MTweb set.
- The top 1,400 topic phrases were mixed randomly with the general phrases until 14,000 queries were obtained.
- These queries were used to collect approximately 550M of CHILweb and approximately 700M words of MTweb data.

Language Model Perplexity and OOV Rate

- Only the first 1000 queries (150M words) from CHILweb was used in addition to the corpora mentioned above.
- Subsets were selected by skipping data from less useful queries, based on their perplexity with an in domain LM built on CHIL and proceedings data.
- The threshold was selected so that each subset contained around 150M words.
- The perplexity on the RT05 eval set was 130.
- The final vocabulary was 52,000 words which yielded a 0.65% OOV rate on the RT05 eval set.
- Addition of the web data to the LM training set reduced WER by 1.2% absolute on the RT05 eval IHM condition.

Speech Features for IHM Segmentation

- Speech activity features are extracted on a per frame basis, with a frame size of 32 msec and a frame shift of 10 msec.
- Speech features for segmentation consisted of:
  - the frame energy in dB,
  - the mean and variance normalized energy passed through a sigmoid function,
  - the energy-normalized linear prediction error,
  - the spectral slope of a mel-warped filter-bank spectrum along the frequency axis,
  - the speech class posterior computed from a multi-layer perceptron (MLP) trained with standard MFCC features to classify speech and non-speech.

IHM Segmentation and Speaker Clustering

Segmentation for the IHM Condition was conducted in three steps:

1. **Background speech activity rejection:**
   - Choose the microphone with the highest energy for each frame.
   - Prune out unreliable microphone switches based on a minimal duration of voiced speech determined from normalized energy, energy-normalized linear prediction error, and speech class posterior from the MLP.
2. **Foreground speech activity detection:** Frames with negative spectral slope, high normalized energy, and low energy-normalized linear prediction error are further tagged as foreground speech. These estimates are further smoothed with a median filter of 0.5 sec duration.
3. **Sentence breaking:** A sentence break is made at the point of highest confidence non-speech (based on lowest average energy level and longest duration) in the interval between 0.5 sec and 15 sec from the current starting point.
4. All details concerning the segmentation algorithm can be found in [Fügen2006b].

Decoding Strategy

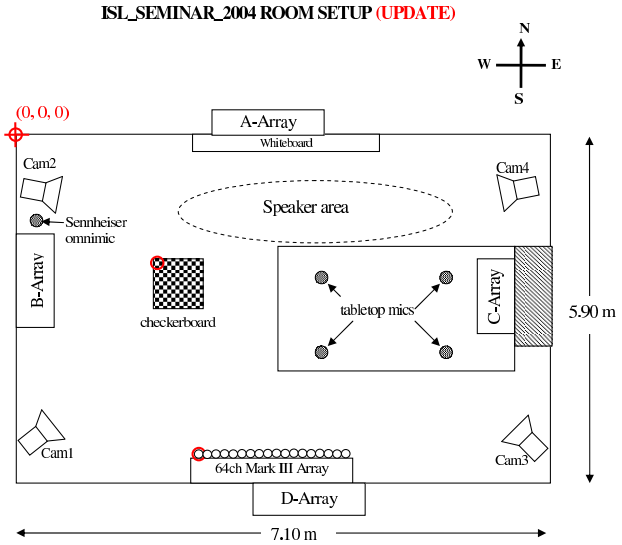
We investigated decoding strategies based on SI, VTLN, and ML-SAT models:

A. adaptation was strictly step-by-step and used only matching models;
B. V+F+M adaptation was applied to all passes;
C. a slight modification of B: only V+M adaptation after first pass;
D. only the speaker-adapted VTLN and ML-SAT models were used: incremental V+F adaptation on first pass.

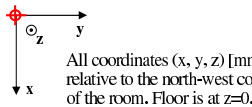
	1st (FFT)	2nd (MVDR)	3rd (FFT)	4th (MVDR)
A	34.2%	30.0%	27.9%	25.5%
B	34.2%	27.0%	25.4%	
C	34.2%	26.8%	25.3%	
D	31.5%	26.5%	25.4%	25.0%

Table 3: Adaptation experiments, with different acoustic models on the IHM condition of Dev.

Sensor Configuration at the University of Karlsruhe



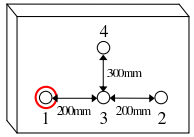
	x	y	z
Checkerboard 2004_11	2130	3260	732
Checkerboard 2004_06/07/08	2000	3110	730
Mark III	5665	2900	1710
Array A1	105	3060	2370
Array B1	2150	105	2290
Array C1	2700	6210	2190
Array D1	5795	4280	2400



All coordinates (x, y, z) [mm] are relative to the north-west corner of the room. Floor is at z=0.

- Mark III: 64 ch, 20mm mic distance
- Checkerboard square size: 105mm. Position of the first *inner* crossing is given.
- Checkerboard for *internal* calibration: 42mm square size
- Room height: 3m
- Camera height: ~ 2.7m

A/B/C/D-Array Layout:



MDM Condition

- The frontends, AM and LM training for the MDM condition was identical to the IHM condition.
- After segmentation, an agglomerative clustering procedure based on a BIC criterion was used to determine which segments were uttered by the same speaker.
- We assume the speech on all microphones is correlated while at least some of the noise is uncorrelated.
- Hence, we can simply sum up all channels pre-shifted by their relative estimated *time delays of arrival* (TDOA) and divide by the number of channels N to attenuate the noise.

Time Delay of Arrival Estimation

- To estimate the TDOA, we calculate the *generalized cross correlation* (GCC)

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G_{12}(\omega)\, e^{j\omega\tau}\, d\omega \tag{1}$$

where

$$G_{12}(\omega) = \frac{X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})|} \tag{2}$$

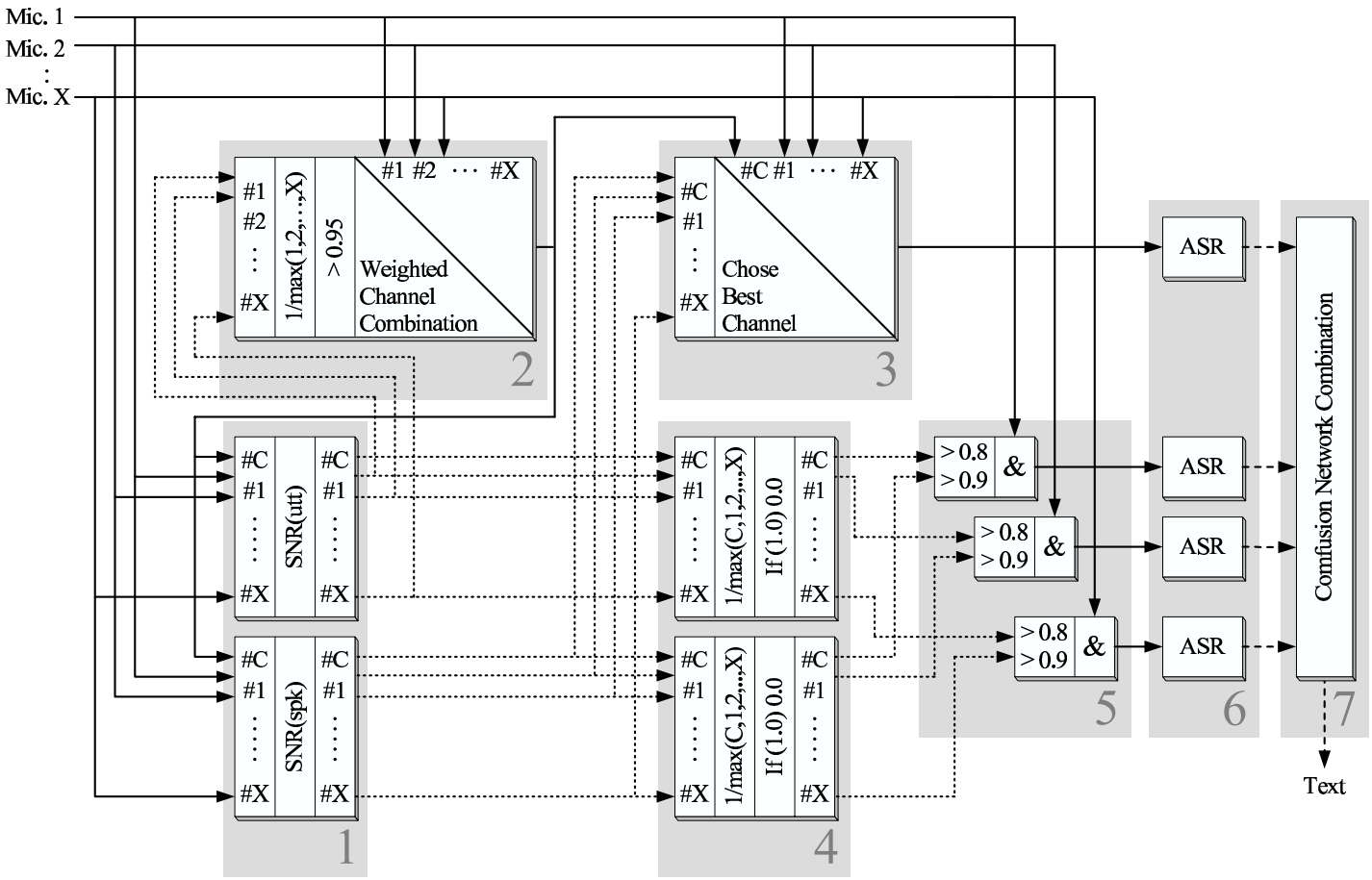
- The estimated TDOA is then given by $\hat{\tau} = \max_\tau R_{12}(\tau)$.
- Let $N_1(e^{j\omega\tau})$ and $N_2(e^{j\omega\tau})$ denote the noise spectral estimates of each channel when no speech is present.
- To improve the TDOA estimate in the presence of correlated noise, $G_{12}(\omega)$ can be replaced with

$$G'_{12}(\omega) = G_{12}(\omega) - \frac{N_1(e^{j\omega\tau}) N_2^*(e^{j\omega\tau})}{|N_1(e^{j\omega\tau}) N_2^*(e^{j\omega\tau})|} \tag{3}$$

Channel Selection and Combination

- A combination of *blind channel combination* (BCC) and *confusion network combination* (CNC) was used for the MDM condition.
- Those channels with a high signal-to-noise ratio (SNR) for a given utterance were combined with BCC.
- This combination produced a 2 dB improvement in SNR and a 4% reduction in WER with respect to the SDM condition.
- A SNR criterion was also used to select which channels were decoded for CNC.
- CNC yielded another 0.5% reduction in WER; see [Wölfel2006] for further details.

Channel Selection and Combination Schematic



Results on Eval 2005 Development Set

- For all conditions, each decoding pass used both an FFT and MVDR model.
- Cross adaptation on the MVDR (FFT) model was performed using the confidence-weighted hypos from FFT (MVDR) model in the prior pass.
- For MDM and SDM conditions, the MVDR alternate phone set model was used only in the last pass in addition to the standard FFT and MVDR models.
- For IHM, the alternate phone set model was used instead of the standard MVDR model in the fourth and fifth passes.

Pass	IHM	SDM	MDM
1st pass	30.3	50.9	46.9
2nd pass	25.0	45.9	42.0
3rd pass	23.9	43.4	38.5
4th pass	23.2		
5th pass	22.9		
RTx	190	110	120

Table 4: Overall system results and real-time factors on Dev.

Mark III Condition

- Source localization was performed based on the output of the T-shaped arrays.
- The speaker tracking algorithm was based on the joint probabilistic data association filter (JPDAF); see [Gehrig2006].
- Using the automatic speaker position estimates, beamforming was performed on the output of the 64 channel Mark III.
- STT was performed on the beamformed output of the Mark III.

Test Set	% Word Error Rate		
	Single Channel	IEKF	JPDAF
RT06 Dev	61.8	49.4	48.8
RT06 Eval	N/A	67.3	66.0

Table 5: STT performance for single channel and beamformed array output using IEKF and JPDAF position estimates.

Comparison of IHM Results on Eval 2006

Pass	IHM	IHM Manual
1st pass	55.2	33.4
2nd pass	50.8	29.4
3rd pass	49.0	28.3
4th pass	47.8	27.3
5th pass	47.1	26.8
Sub/Del/Ins	16.3/10.2/20.6	16.2/6.7/4.0

Table 6: IHM results for each decoding pass on the evaluation 2006 set.

Site	AIT	UKA	IBM	ITC	UPC	Non-Interactive	Interactive
Automatic	65.2	43.2	42.5	38.1	56.7	48.0	43.9
Intersegment	32.0	17.0	16.0	5.0	10.0	N/A	N/A
Manual	31.3	23.8	24.4	30.2	35.1	27.2	25.4

Table 7: Per site IHM results for manual and automatic segmentations on the eval 2006.

Conference Meeting Task

- Acoustic models and decoding strategy were the same used in the lectmtg task.
- Separate language models for the AMI and other data were trained.
- Additional data was harvested from the web for the AMI LM.
- OOV rate was 0.48% (0.57%) on the RT05 (RT06) development set.

LM Set	Eval05	Eval06
AMI	95	95
Other	91	96

Table 8: LM perplexities on conference meeting task.

Conference Meeting IHM Results

Pass Set	Eval05	Eval06
2nd pass	35.2	31.9
3rd pass	33.7	30.8
4th pass	32.6	30.2
5th pass	31.9	30.2
Sub/Del/Ins	N/A	13.1/13.8/3.3

Table 9: IHM results on the evaluation 2006 set.

References I

[Fiscus2006] Jonathan Fiscus, Jerome Ajot, Nicolas Radde, and Christophe Laprun, "Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech," in Proc. LREC, 2006.

[Fügen2006a] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stüker, S. Vogel, and A. Waibel, "Open Domain Speech Recognition and Translation: Lectures and Speeches," in Proc. ICASSP, 2006.

[Fügen2006b] Christian Fügen, Matthias Wölfel, John W. McDonough, Shajith Ikbal, Florian Kraft, Kornel Laskowski, Mari Ostendorf, Sebastian Stüker, Kenichi Kumatani, "Advances in Lecture Recognition: Interactive System Laboratory's RT-06S Evaluation System," in Proc. Interspeech, submitted for publication, 2006.

References II

[Gehrig2006] Tobias Gehrig and John McDonough, "Tracking Multiple Simultaneous Speakers with Probabilistic Data Association Filters," in Proc. MLMI, May, 2006.

[McDonough2002] J. McDonough, T. Schaaf and A. Waibel, "On Maximum Mutual Information Speaker-Adapted Training," in Proc. ICASSP, 2002.

[Wölfel2005] Matthias Wölfel and John W. McDonough, "Minimum Variance Distortionless Response Spectral Estimation: Review and Refinements," IEEE Signal Processing Magazine, September, 2005.

[Wölfel2006] Matthias Wölfel, Christian Fügen, Shajith Ikbal, and John W. McDonough, "Multi-Source Far-Distance Microphone Selection and Combination for Automatic Transcription of Lectures," in Proc. Interspeech, submitted for publication, 2006.